DOCUMENT RESUME

ED 035 588 24 SP 003 446

AUTHOR Lauroesch, William P.; And Others
TITLE The Use of Student Feedback in Teacher Training.
INSTITUTION Chicago Univ., Ill. Graduate School of Education.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-8-E-115
PUB DATE Jun 69
GRANT OEG-5-9-235115-0001(010)
NOTE 56p.

EDRS PRICE EDRS Price MF-$0.25 HC-$2.90
DESCRIPTORS *Feedback, *Student Opinion, *Teacher Education, *Teacher Evaluation, Teacher Interns, Teacher Supervision
IDENTIFIERS SOQ, Student Opinion Questionnaire

ABSTRACT

A study was conducted to evaluate certain formats for providing feedback of student opinion during teacher training. A class of 69 interns were randomly assigned to five treatment groups and one of their classes was chosen to be used in the experiment. In Febraury each class in three of the groups was given the Bryan Student Opinion Questionnaire (SOQ) and a parallel self-appraisal form only, and a fifth group was not tested until the end of the experiment. The three groups of teachers whose pupils were given the SOQ were given three feedback treatments; (1) no feedback, (2) written feedback only, and (3) written feedback plus a conference with a supervisor. In early May all five groups were given the SOQ and the self-appraisal form. The data were analyzed in three stages: (1) to compute basic statistics, correlations, and factor analyses, (2) to detect differences between treatment groups, and (3) to detect possible sources of invalidity and to assess the accuracy of self-appraisal. There were significant differences between the treatment groups. Feedback, as compared with no feedback, was effective in changing subsequent pupil ratings, but in a direction opposite that which was expected. Those teachers receiving feedback were rated lower in May than in February. The conference mitigated this effect, but did not eliminate it. (Findings, instruments, and references are included.) (JS)

ED035588

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FINAL REPORT

Project No. 8-E-115

Contract No. OEG-5-9-235115-0001(010)

THE USE OF STUDENT FEEDBACK IN TEACHER TRAINING

William P. Lauroesch

Peter D. Pereira

Kevin A. Ryan

June 1969

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

SP003446

Master of Arts in Teaching Program
Graduate School of Education
University of Chicago
Chicago, Illinois

TABLE OF CONTENTS

PREFACE v

SUMMARY 1

INTRODUCTION. 4

METHOD. 9

DATA ANALYSIS 12

RESULTS 19

DISCUSSION. 30

REFERENCES. 33

APPENDIXES

A. Student Opinion Questionnaire
B. Self-Appraisal Questionnaire
C. Sample of Written Feedback
D. Protocol for Feedback Conferences
E. Statistical Tables

LIST OF TABLES

Table		Page
1.	Rank Order of Ratings on the Pupil Questionnaire.	20
2.	Mean Total Change Scores for Three Treatment Groups	23
3.	Mean Vectors of Gain Scores for Three Treatments	24
4.	Chi-square Analysis of the Pattern of Significant Gains and Losses.	25
5.	Mean Deviation and Total Number of Students Tested	27
6.	Correlations Between Self-Appraisal and Pupil Questionnaire.	28
7.	Gains and Losses in Accuracy of Self-Appraisal for Three Treatments	29
8.	Basic Statistics for Class Means Pupil Questionnaire All Pre-Test Groups.	E-1
9.	Basic Statistics Self-Appraisal Questionnaire All Pre-Test Groups.	E-1
10.	Basic Statistics for Class Means Pupil Questionnaire All Post-test Groups	E-2
11.	Basic Statistics Self-Appraisal Questionnaire All Post-test Groups	E-2
12.	Intercorrelations of Class Means for the Pupil Questionnaire.	E-3
13.	Principal Components Factor Loadings	E-4

14. Intercorrelations of Ratings on the Self-Appraisal Questionnaire. E-5

15. Summary of t-test comparisons Control Group E-6

Summary of t-test comparisons Experimental Group I. E-7

Summary of t-test comparisons Experimental Group II E-8

PREFACE

This study was conceived in an attempt to evolve efficient, as well as effective, ways to help teachers use their students' opinions about them. It continues earlier work by the principal investigator, Kevin Ryan, who had long been impressed with the potential of student feedback. The other two investigators, Bill Lauroesch and Peter Pereira, brought to the project a total of thirty years of teaching experience during which time they had had many occasions to observe the perceptiveness of their students.

Many people have assisted us in this project. We owe special thanks to Dr. Roy C. Bryan who has done research in the area of student feedback for over thirty years. This study was substantially advanced by his research and personal assistance, and he was most generous in sharing his time and materials. Roald Campbell, Dean of the Graduate School of Education of the University of Chicago, gave us strong encouragement and support.

This study would not have been possible without the generous cooperation of the MAT interns of the University of Chicago. Our work was greatly facilitated by their willingness to rearrange their schedules in order to allow us enough time to administer the questionnaire.

Many people have contributed in special ways. Our thanks are due to Don Kolakowski, who helped us with the data analysis, and to Elizabeth Hodge, who oversaw the day-to-day activities of the study with unusual grace.

We especially wish to express appreciation for the service and patience of Estelle Buccino, who had the difficult job of deciphering and typing the students' comments, and Vicky Leak, who has prepared the final report.

SUMMARY

Thirty years of research and development have indicated that student feedback is a useful and reliable means for informing and directing behavior l changes in teachers. The research has not shown that student feedback can appreciably effect teacher behavior within the time boundaries of practice teaching or internship experience. The general goal of this study was to evolve efficient and effective means for utilizing student feedback in teacher education. Is written feedback by itself sufficient to produce change in behavior? Or must this information be supplemented by a conference which directs the student teacher or intern in how to use the feedback? Will teachers be able to accurately predict how their students will rate them? This study was designed to answer these, and other, questions.

The entire class of MAT interns at the University of Chicago were randomly assigned to five treatment groups and one of their classes was chosen to be used in this experiment. In February, 1968, each class in three of the groups was given the Bryan "Student Opinion Questionnaire" (SOQ) and a parallel self-appraisal form. A fourth group of teachers was given the self-appraisal form only, and a fifth group was not tested until the end of the experimental period. The three groups of teachers whose pupils were given the SOQ were given three feedback treatments as follows: (1) no feedback; (2) written feedback only; and (3) written feedback plus a conference with a supervisor. In early May, eight school weeks after the first testing, all five groups were given the SOQ and the self-appraisal form.

The data were analyzed in three stages: (1) to compute basic statistics, correlations, and factor analyses; (2) to detect differences between treatment groups; and (3) to detect possible sources of invalidity and to assess the accuracy of self-appraisal. The results of stage one of the analysis indicated that the items of the SOQ are highly correlated with a tightly clustered factor structure. The means on the items are towards the middle of the range leaving room for improvement, or deterioration.

Stage two of the data analysis showed that there were significant differences between the treatment groups. Feedback, as compared with no feedback, was effective in changing subsequent pupil ratings, but in a direction opposite to that which was expected. Those teachers

receiving feedback were rated lower in May than in February. The conference mitigated this effect, but did not eliminate it. Evidently, over a short period of time, the information provided by feedback results in a sharp drop in performance although the direction given in the conference is helpful in reducing this effect.

Stage three of the data analysis did not detect any sources of invalidity due to a tendency for ratings to change over time or due to reactive effects. The self-appraisal was accurate for items on which there can be said to be an objective answer (e.g. knowledge of subject matter) but no better than chance on items where the student's opinion has to be accepted (e.g. clarity of explanations). The group which received written feedback plus a conference was significantly more accurate in predicting in May than in February, indicating that the conference helped teachers to understand the meaning of the feedback.

The results of this study demonstrate the power of student feedback, but they also show that information alone is likely to be more damaging than helpful. Further research is needed to develop more effective ways of using feedback so that beginning teachers can improve their performance over a short period of time.

THE USE OF STUDENT FEEDBACK

IN TEACHER TRAINING

INTRODUCTION

THE PROBLEM

Because the pupils in a student teachers' class have an opportunity to observe all of his classroom behavior, they are in a position to be the most knowledgeable source of information about his day-to-day classroom performance. This does not mean that they are the most discriminating or most perceptive sources of information about teacher performance; indeed, it would be unreasonable to expect pupils to be expert judges of all aspects of a teaching performance. Yet, collectively there may be some aspects of teaching which students are able to observe quite reliably. If one were able to tap this source of information and have confidence in it, it would be a valuable supplement to the information gained from other sources. The purpose of this study, then, was to evaluate certain formats for providing feedback of student opinion during teacher training.

OBJECTIVES

The general objectives of this study were:

1. To establish that significant changes in student teacher behavior can be produced in six to eight weeks by the use of feedback of pupil opinion combined with a self appraisal by the student teacher.

2. To compare the effectiveness of two different methods of providing feedback: conference with their supervisor and written report.

3. To determine the extent to which the administration of the pupil rating form and the self-appraisal form produced changes in behavior when the teacher received no feedback.

4. To determine how accurately student teachers can predict their pupils' opinions and whether feedback from pupils will improve the accuracy of predictions.

RELATED RESEARCH

A study at Stanford in the spring of 1965 indicates the potential of feedback of student opinion in teacher education.[1] During the year

---

[1] H. E. Aubertine, "An Experiment in the Set Induction Process and its Application in Teaching" (unpublished Ph. D. dissertation, Stanford University, 1965).

of training, intern teachers were given feedback from three sources: their university supervisor, their resident supervisor in the school, and their pupils. Feedback from the university supervisor was given by means of conferences and written reports after classroom observations. Each intern had an average of 18 conference-reports from his university supervisor. Feedback from the resident supervisor followed the same conference-report format, each intern having an average of 31 such conference-reports. The student feedback was given in two ways: (1) a summary of numerical ratings of the intern teacher on a 13-item teacher competency appraisal guide, and (2) a typed verbatim transcript of the pupils' responses to three general questions about their teacher's strengths and weaknesses. Student feedback was given to the interns only twice. At the end of the year the interns were asked, "Generally which source of feedback (from university supervisor, resident supervisor, or pupils) has been most helpful to you?" The highest percentage of interns perceived that student feedback had been most helpful, even though they had received it only twice. Further, 98 per cent claimed that the typed sheets of students' comments were more helpful than the summary of numerical ratings on the teacher competency appraisal form. It would appear from this brief study that the feedback easiest to obtain (viz., student feedback) was the most valued.

In a subsequent study at Stanford Ryan pursued further the question of student feedback.[2] He considered different methods of presenting the feedback to determine whether the feedback was more powerful when a supervisor helped the intern to interpret it or if it was sufficient for the intern to read the feedback for himself. He also wanted to find out which teachers were most receptive to feedback. Unfortunately, his study did not give clear answers because none of the hypotheses were supported by statistically significant data. For this he suggests reasons which are important to keep in mind when designing further studies. Most of the reasons are criticisms of the instrument used to obtain student opinion. It was subject to the "halo effect," that is, the tendency for a pupil to respond with a total acceptance or rejection of his teacher's behavior. The pupils were also inclined to be so complimentary to their teachers that it was difficult to see what needed improvement. The instrument's reliability was not checked. Ryan suggests that a feedback instrument ask students to rate their teacher's skills on a scale and then ask for comments on how he can improve particular skills.

The most careful development of a reliable procedure for soliciting student opinion has been done by Bryan over a period of thirty years. His questionnaire follows much the pattern as suggested by Ryan. The reliability of each scale has been checked frequently, those scales with low reliabilities having been replaced by scales with higher reliabilities. Generally, the reliability coefficients range from .80 to .90. Inter-correlations between scales are considerably lower, indicating that

---

[2]Kevin A. Ryan, "The Use of Students' Written Feedback in Changing the Behavior of Beginning Secondary School Teachers" (unpublished Ph. D. dissertation, Stanford University, 1966).

there is a reasonable freedom from the "halo effect." Although average responses on all scales tend to be quite favorable to a teacher, there is usually room for improvement. The scales are particularly useful in showing areas of relative strength and weakness since one teacher will quite often be placed in quite different positions on the separate scales. Each student is asked for specific comments, and his attention is directed to the scales where he has indicated that his teacher has a weakness.

Bryan has made some careful checks on whether his instrument is effective in changing teacher behavior. In a study done over a period of two years (1960-62), he measured the change in responses on ten scales.[3] One group received no feedback of student opinion. Another group received feedback twice by means of a written report which was mailed to them. In both groups there were teachers of varying years of experience, teachers in all academic areas, and teachers from large and small schools. The gains and losses of each teacher on each scale were checked for statistical significance at the .01 level. The experimental teachers made considerably more significant gains on every scale than the control teachers. Conversely, experimental teachers had consierably fewer losses on all but one scale. Bryan has checked his data for alternative explanations, but it is hard to escape his conclusion that feedback of student opinion can help many teachers change their behavior, at least as perceived by their students.

Bryan's questionnaire has been used in a more recent study by Tuckman and Oliver.[4] Students in high school vocational subjects were asked to rate their teachers at the beginning and the end of a twelve-week period. The teachers were randomly divided into four treatment groups. One group received no feedback. The others received feedback from their students only, from their supervisors only, or from both their supervisors and their students. It was found that student feedback led to a positive change in pupil ratings, as compared with no feedback, while the supervisor feedback led to a negative change. When combined with student feedback, supervisor feedback had little additional effect. Thus Bryan's findings are replicating over a shorter period of time, and, in addition, student feedback appears to be more influential than supervisor feedback.

Several features of Tuckman and Oliver's study should be noted. First, no significant relationship was found between years of experience and receptivity to feedback, although the most experienced group tended to be less receptive to feedback. Secondly, an additional group of teachers was rated by students at the end of the twelve-week period only.

---

[3]Roy C. Bryan, Reactions to Teachers by Students, Parents and Administrators, Report of Cooperative Research Project No. 668, U. S. Office of Education (Kalamazoo, Michigan: Western Michigan University, 1963).

[4]B. W. Tuckman and W. F. Oliver, "Effectiveness of Feedback to Teachers as a Function of Source," Journal of Educational Psychology, LIX (1968), 297-301.

This post-test-only group was not significantly different from the no-feedback group. Thus there is no indication that the prior administration of the pupil rating form by itself had any influence on subsequent rating. Thirdly, all the change scores are negative, i.e., all groups rated their teachers more harshly towards the end of the year than in the middle, the changes in student feedback group being significantly less negative. Tuckman and Oliver explain this by saying, "At the time when the teacher is about to evaluate and grade the student, the student perhaps replies in kind."[5]

A rationale explaining why teachers are likely to change their classroom behavior when they are provided with information on their pupils' opinions has been developed by Gage, Runkel, and Chatterjee.[6] This rationale is based on the premise that the feedback will create an imbalance that the teacher will move to correct. His most likely response would be to modify his behavior or at least modify students' perceptions of his behavior, although other reactions are possible. For instance, he might distort his own perceptions of the feedback to rationalize a more palatable picture.

Gage, Runkel and Chatterjee developed their rationale to support a study on feedback. They wanted to know whether teachers who received feedback of pupils' opinions would modify their behavior more than teachers who received no feedback. They found that the feedback not only produced change in behavior, but also increased a teacher's self-awareness in the sense that he was better able to predict his students' opinions.

Two aspects of Gage's work are particularly important for this study. First, although he found statistically significant differences, the differences themselves were not large. Only sophisticated data analysis was able to detect their significance, a result which casts doubt on the efficiency of feedback. Gage suggests that the reason for such small differences is that his procedure for administering feedback was not specifically designed to have maximum impact. Subsequent studies, he suggested, should consider more effective ways of administering feedback. Secondly, the process of administering the feedback instruments to students (and a corresponding one to teachers) appeared to produce a change in behavior even without feedback. If there had been a "post-test-only" group, it would have helped to detect this reactive effect.

To summarize, a reliable and useful instrument exists for measuring pupils' opinions of their teachers. This has been used to change teacher behavior (as perceived by their pupils) in at least two experiments. No

---

[5] Ibid., p. 300-301.

[6] N. L. Gage, P. J. Runkel, and B. B. Chatterjee, Equilibrium Theory and Behavior Change: An Experiment in Feedback from Pupils to Teachers. (Urbana, Illinois: University of Illinois Bureau of Educational Research, 1960).

study has been reported in which significant behavior change occurred with student teachers over a short time span (six to ten weeks) as a result of feedback of student opinion. If feedback is to be effective over a short period of time, it should be administered so as to have maximum possible impact. There is some evidence to indicate that students get harsher in their judgment as the year progresses. There is contradictory evidence of whether the process of self-evaluation and/or evaluation by pupils is sufficient to produce behavior change even without feedback. It is reasonable to believe that some behaviors are more readily changed than others, that some teachers are more susceptible to feedback than others, and that some school climates foster change more easily than others. None of the research which has been done has been designed to answer these questions.

METHOD

GENERAL DESIGN

In late February, 1968, pupils of student teachers were asked to rate their teachers by using the current version of the questionnaire developed by Bryan. (See Appendix A.) At the same time each teacher was asked to appraise himself by means of an adaptation of the pupil questionnaire. The items were the same, but each teacher was asked to respond by predicting the way in which his students would answer the questions. (See Appendix B.)

Two groups of student teachers were provided with feedback. One group was mailed a written summary of the information collected from his pupils arranged so that he could compare it with his own predictions. (See Appendix C.) The other group received the written summary during a conference with a supervisor. Together the student teacher and the supervisor considered the areas of relative weakness and the place where the student teacher was least successful in predicting his students' responses. They then decided on two areas which particularly needed improvement. Specific suggestions for improvement were agreed upon. During the next eight weeks, the student teacher was expected to concentrate on improving his pupils' opinions of him in these areas, with the knowledge that his pupils would be evaluating him again. Every attempt was made to insure that the format of the conferences was uniform. A protocol for these conferences is included in Appendix D.

POPULATION AND SAMPLE

All teaching interns from The University of Chicago MAT Program who were teaching in high schools in the Chicago area were randomly assigned to five groups: a control group, two experimental groups, a self-appraisal group, and a post-test-only group. Then one of the classes which they were teaching was randomly chosen to be used in this experiment.

The MAT interns were in the second year of a two year teacher preparation program. As teaching interns, they were regularly scheduled classroom teachers with all the accompanying responsibilities and frustrations; but most of them taught only three fifths of a normal teaching load and were therefore paid proportionally. Almost all of the interns had come to the MAT program directly from liberal arts colleges. During their first year in the program, prior to the internship year, they combined work in education and graduate study in their teaching field.

The study took place in thirty-one high schools in the metropolitan area of Chicago. Five of these were private schools and the rest public. Both the public and private schools, however, serve a cross section of the secondary school population. Seven of the schools were suburban high schools catering to an upper middle class, white community. On the other hand, seven were urban secondary schools located in poor, all black communities. The remainder of the schools serve communities somewhere between these extremes.

TREATMENTS

The control group (16 teachers) was given a self-appraisal form and their pupils were given rating forms. No feedback was given to these teachers until after the experimental period of eight weeks. At the end of the experiment the same instruments were administered a second time.

The experimental groups (16 teachers each) were given the self-appraisal form and their pupils were given rating forms. These teachers received feedback in the two ways described above. At the end of the experimental period the same instruments were administered again.

The self-appraisal group (10 teachers) was given the self-appraisal form at the beginning of the experiment, but their pupils were not given rating forms. At the end of the experiment, the teachers were given a self-appraisal form and their pupils were given rating forms.

The post-test-only group (11 teachers) was given rating forms at the end of the experimental period. They were not given any other tests.

Diagramatically, the design looks like this:

Control Group (16)	Self-Appraisal	Pupil Questionnaire		Self-Appraisal	Pupil Questionnaire
Experimental Group I (16)	Self-Appraisal	Pupil Questionnaire	Written Feedback	Self-Appraisal	Pupil Questionnaire
Experimental Group II (16)	Self-Appraisal	Pupil Questionnaire	Written Feedback & Conference	Self-Appraisal	Pupil Questionnaire
Self-Appraisal Group (10)	Self-Appraisal			Self-Appraisal	Pupil Questionnaire
Post-Test Only Group (11)				Self-Appraisal	Pupil Questionnaire

Three teachers were dropped from the experiment because we were unable to administer the second test to their classes. One teacher was dropped from the control group because of a prolonged illness, and two were dropped from the second experimental group because conditions in their schools made it impossible to administer the test a second time.

DATA ANALYSIS

The data were analysed in three distinct phases. (1) We computed basic statistics and correlations together with a factor analysis of two of the correlation matrices. (2) We analysed the differences between treatment groups. (3) We considered the variety of opinion within a class and the accuracy of the teachers' self-appraisal.

Phase two, comparisons between treatment groups, required extensive analysis. In order to reflect the complexity of the data, we felt it important to consider two sources of variation which are sometimes overlooked: (a) Variation due to differential effects of the items; i.e. not all items may be equally effective in distinguishing between groups. (b) Variation due to individual differences between students within classes. There may be wide disagreement about one teacher within his class while students in another class may be quite uniform in their judgements.

In order to deal with this complexity, we compared treatment groups in three ways. First, we used a total gain score following an approach similar to that of Tuckman and Oliver. We then used a multivariate approach in order to gain a better understanding of what each item contributed to the differences between groups. Finally, we performed multiple t-tests, following Bryan's approach, in an attempt to reflect some of the effects of within-class variation in our analysis.

The analyses of the data are described in detail in this section of the report. The results of the analyses are presented in the next section.

BASIC STATISTICS

The responses to both the pupil questionnaire and the self-appraisal questionnaire were assigned weights from one to five, the lowest score with which a teacher could be rated being 1.00 and the highest being 5.00. (See the sample graph in Appendix C.) Means and standard deviations on the pupil questionnaires were then computed for each class. The individual responses of the teacher to the self-appraisal questionnaire were simply recorded. These statistics were then used to prepare the written feedback to individual teachers. All subsequent calculations on pupil feedback were based on class means. Thus, the measures for pupil feedback can be thought of as continuous variables ranging from 1.00 to 5.00. The self-appraisal measures, however, must definitely be thought of as discrete variables since they can only take on the values 1, 2, 3, 4, or 5.

The class means were then used to compute basic statistics (mean, standard deviation, variance, skewness, and kurtosis) for each group separately, for all pre-test groups combined, and for all post-test groups combined.

CORRELATION AND FACTOR ANALYSIS

Using class means, two correlation matrices were produced for all pre-test groups combined, for all post-test groups combined, and for some groups separately. In each case the two matrices were: correlations between scores on the pupil questionnaire and correlations between scores on the self-appraisal questionnaire.

Two of the correlation matrices for the pupil questionnaire were factor analyzed, using Hotelling's principal components solution, and then rotated, using an orthogonal varimax technique which simplifies the columns of the factor matrices.

ANALYSIS OF GROUP DIFFERENCES USING A TOTAL GAIN SCORE

Tuckman and Oliver's data were similar to ours. They used a comparable form of Bryan's questionnaire with ten items, they had one control group and three experimental groups, and they used a short experimental period (12 weeks). In order to compare groups, they computed a single gain score as follows:

"The measure of change in each condition was the sum of the differences between the pre-interval judgements by the students on the 10 items and their post-interval judgements. Ratings on each item were averaged across students and the pre-interval average on each item was then subtracted from the post-interval average to yield a change score on each of the 10 items. These 10 item change scores were summed to obtain a total change score."[7]

We computed a total gain score in the same way except that our questionnaire has 12 items. (The item on homework is scaled differently, and therefore it was not used in making group comparisons.) This procedure reduces a multi-variate problem to a univariate problem by adding together highly correlated variables. This is not invalid, but it does result in a considerable loss of information because it pools together all the knowledge which we have about individual items. Thus it may cover up differences between groups when, for example, the total change scores are the same but the items on which changes are made are different. On the other hand, as we shall see, it may reveal differences between groups more clearly when the information which is lost is confusing or contradictory.

---

[7]Bruce W. Tuckman and Wilmot F. Oliver, "Effectiveness of Feedback to Teachers as a Function of Source," Journal of Educational Psychology, LIX (1968), 299.

Total gain scores were computed for the three groups which were given the pupil rating form twice; i.e. for the control group and for the two experimental groups. These groups were then compared by using a one-way, univariate analysis of variance with Helmert contrasts. The purpose of using Helmert contrasts was, first, to contrast the two experimental groups to see if the different forms of feedback had different effects. Then, if the experimental groups were judged to be similar, to contrast the control group with the two experimental groups pooled together. These contrasts are in line with the general rationale of the study; namely, that feedback has an effect compared with no feedback and that there are differential effects due to different ways of administering feedback.

The computations for this analysis and for all subsequent analyses of variance, multivariate and univariate, were done on the IBM 7094 at the University of Chicago Computation Center using the current version of MESA 97. The statistical basis of this program has been explained by Bock,[8] and the program itself has been described by Finn.[9]

ANALYSIS OF GROUP DIFFERENCES USING MULTIVARIATE ANALYSIS OF VARIANCE

As we have mentioned, the univariate analysis results in a loss of information because it reduces twelve variables to a single variable. The problem which we wanted the data analysis to attack was to detect and characterize differences among the three groups on the twelve variables simultaneously. Multivariate analysis is particularly appropriate for this purpose.[10]

Twelve gain scores were computed for each teacher by subtracting his pre-test scores from his post-test scores. These were then treated as dependent variables in a one-way, multivariate analysis of variance to compare the control group with the two experimental groups. Helmert contrasts were again used for the same reasons as in the univariate analysis. Variables were entered into the analysis in the order of importance which we attached to them. This admittedly arbitrary order was: control, interest, variety, planning, clarity of explanation, fairness, knowledge of subject matter, attitude toward students, student participation, sense of humor, attitude toward subject, attitude toward student opinions.

---

[8]R. D. Bock, "Programming Univariate and Multivariate Analysis of Variance," Technometrics, V (February, 1963), 95-116.

[9]J. D. Finn, "Univariate and Multivariate Analysis of Variance and Covariance," Research Memorandum No. 3, Statistical Laboratory, Dept. of Education The University of Chicago, April, 1966.

[10]R. D. Bock and E. A. Haggard, "The Use of Multivariate Analysis of Variance in Behavioral Research," in Handbook of Measurement in Education, Psychology and Sociology, edited by Dean Whitla (Boston: Addison, Wesley, 1968).

There were two reasons to compare all five groups: (1) to detect any tendency for class means to change over time; and (2) to check to see whether the process of rating oneself and being rated by one's students was by itself effective in bringing about change. In order to do this, we used class means on the twelve items of the pupil questionnaire as dependent variables in a one-way, multivariate analysis of variance using simple contrasts. The simple contrasts were used to compare each group separately with the post-test only group. Variables were entered into the analysis in the same order as before.

We were also attracted by the idea of using multivariate analysis of covariance to compare the three groups, using the twelve pre-test scores as covariates and the twelve post-test scores as dependent variables. The rationale for this is that one's level in one area influences, favorably or unfavorably, one's subsequent performance in all areas. For example, a person who is initially low in control may find it hard to raise his score without lowering other scores. This inter-relationship may be intensified for a group which received feedback because the knowledge of what students think about one's performance in a single area may influence subsequent behavior in all areas. Thus covariance analysis might give an indication of the dynamic quality of feedback. In fact, when we tried this analysis, it gave us no information not already provided by the multivariate analysis of gain scores.

ANALYSIS OF GROUP DIFFERENCES USING T-TESTS ON INDIVIDUAL STATISTICS

Bryan[11] analyzed the differences between his treatment groups by using t-tests. The analyses described in this section follow the same general plan, but they extend Bryan's approach somewhat further. This analysis takes account of within-class variability to the extent that the significance of a gain or loss depends on the size of the within-class variance.

It is important to notice that the t-tests do not assess the significance of differences between groups. They were intended to assess the significance of a gain or a loss; therefore, they were done individually for each teacher on each item. Thus, since there were 45 teachers and 12 items, we had to make 45 times 12, or 540 separate t-tests. The validity of this approach is open to question since the 12 t-tests performed on the scores for a given teacher are certainly not independent. One way to avoid this difficulty would be to use Hotelling's T^2, an approach which we did not pursue because a preliminary analysis indicated it would be unprofitable with these data.[12]

---

[11]Bryan, op. cit.

[12]P. J. Rulon and W. D. Brooks, "On Statistical Tests of Group Differences," in Handbook of Measurement in Education, Psychology and Sociology, edited by Dean Whitla (Boston: Addison-Wesley, 1968).

The t-tests were done for two tails of a t-distribution. This assumes that the two samples are independent samples from normal populations with the same variance and tests the null hypothesis that the means are the same against the alternative that they are not the same. The assumption that the variances were the same was tested by an F-test. Overall there were not more significant F-ratios than one would expect by chance. There were no important differences between the groups in this respect. Thus there is no reason to question the assumption that the variances were the same in May as they were in February. Nevertheless, in each case where there was a significant t and a significant F, the t-test was refigured using the Welch procedure. In no case did this change the level of significance of the gain or loss. Nor would we expect it to do so, since the class sizes in each case were approximately equal. The t-test is not sensitive to violations of the assumption of equality of variances when sample sizes are close to the same.[13]

Once the t-tests were completed, further analysis was needed in order to see whether there were differences between experimental groups. Bryan's data were clear in this respect: the group which received feedback showed a far higher proportion of significant gains. Our data are less clear, so that additional analysis was needed to test whether the pattern of significant gains or losses was significantly different from chance expectation. This was done using a chi-square test. Chi-square was computed for each group separately since there were three independent comparisons of how well each group fit the distribution predicted under the null hypotheses. The catagories are mutually exclusive and exhaustive, but they have been combined in some cases so that not more than twenty per cent of the expected frequences are less than five. By using chi-square in this situation we assume that the observations are independent, i.e., that for one teacher the significance of gain on one question is independent of the significance of a gain on another question. As with the t-tests, this is a highly questionable assumption, but one which it is hard to avoid making with this approach to the data.

The chi-square analysis does not tell us anything about the direction of change. In order to test this we used sign-tests by which we tested the null hypothesis that the median of the differences was zero against the alternative that the median was non-zero. All of the sign-tests are two tailed tests because we had no reason to believe that scores should gain rather than lose. When N was larger than one hundred, we used a normal approximation.[14]

---

[13] K. A. Brownlee, Statistical Theory and Methodology in Science and Engineering (New York: John Wiley & Sons, Inc., 1960), Chapter 9.

[14] Ibid., Chapter 7.

WITHIN GROUP MEAN SQUARES

Bryan assessed the reliability of his instrument by computing split-half correlation coefficients (and Spearman-Brown correlation coefficients) on all questions using fifty classes to which the questionnaire was administered. This presumably gives an estimate of how much agreement there would be between two groups of pupils observing the same teacher. As reported earlier, Bryan found uniformly high reliabilities.

If one thinks of the teacher as possessing a competence, such as a thorough knowledge of his teaching field or an ability to stimulate interest, then one might think that pupils are judging the extent to which he has or does not have this ability. There is, in a sense, a right answer, and the pupils are trying to give us a reliable estimate of this answer. From this point of view Bryan's method is appropriate.

Yet we could also view the situation in another way. Each student could be telling us about his perception and not about an objective fact to which there is, in theory, a correct answer. From this point of view, what the student reports is the best estimate we have of his perceptions, and no statistical method can tell us about the reliability of this estimate.

An important implication of this point of view is that the average of a whole class, or of half a class, can mask a great deal of disagreement, disagreement which it is important to notice. In order to get some measure of this disagreement, we computed within group mean squares for all pre-test groups combined and for all post-test groups combined. These statistics give us a measure of the variability of opinion about teachers in their classes.

REGRESSION ANALYSIS

One of the objectives of the study was to assess the accuracy of the teachers' self-appraisal. The statistical analysis is complicated by the fact that the self-appraisal scores are discrete variables with a limited range while the class means are continuous variables. This problem could have been avoided had we asked the teachers to rate themselves on a scale with more intervals, but this would also have presented further problems of scaling and response bias.

There are two parts to the question of accuracy. In order to assess the relative accuracy of the self-appraisal, the self-appraisal scores on each item were correlated with the corresponding class means on the pupil questionnaire. Those items on which the correlation coefficients were significantly different from chance expectation were noted.

In order to assess whether there was any change in accuracy of self-appraisal due to feedback, we compared residual variances. As an example, suppose we wished to compare the control group in February with the control group in May on item 1. In February, the correlation between self-appraisal and pupil-rating was .56 which accounts for about thirty per cent of the variation. The best estimate of the residual

variance, that not accounted for by regressions, is .104 with 14 degrees of freedom. Similarly, the best estimate of the residual variance on item 1 in May was .246 with 13 degrees of freedom. The F-ratio to test the null hypothesis that these are estimates of the same variance is 2.44 with 13 and 14 degrees of freedom. This is not quite significant at the .05 level, so we can conclude that there was no change in accuracy on this item for this group.

Of course, this analysis overlooks the inter-relationships between the items. We cannot generalize to a test with different items. If, for instance, we excluded some items, we might get different results. The accuracy of self-appraisal, or lack of it, may be due to the fact that the other items are in the test.

RESULTS

BASIC STATISTICS

Complete tables of the basic statistics for all pre-test groups combined and all post-test groups combined are presented in Appendix E, Tables 8-11. The discussion in this section presents some of the information which can be derived from these tables.

Means--a graph of the means for all pre-test groups combined is presented in Figure 1. The graph for all post-test groups would look substantially the same with respect to both level and shape. Notice that these are averages of class means; therefore, this graph is relatively flat compared to the similar graph for a single teacher which is presented in Appendix C.

With respect to level, these data are similar to those reported by Bryan for 100 first year teachers.[15] Bryan's questionnaire at that time was slightly different from the current version so that not all questions are comparable. Where the questions are the same (i.e. questions 1-6), the MAT population was rated at about the same level on knowledge, fairness, and attitude toward the student. They are somewhat lower in clarity of explanations, and markedly lower in control and the ability to stimulate interest. These differences could be due to differences between our MAT interns and Bryan's sample of first year teachers; but they also could be due to differences between the two samples of pupils who were rating their teachers; or they could be due to some other factor such as the time of year or a tendency for pupils to be more critical now than they were when Bryan's data were collected. Because of this difficulty of interpretation significance levels for these differences were not computed.

With respect to the rank order of questions, the MAT population and Bryan's population are also similar. Control for both groups is rated lowest, the other comparable questions being ranked in about the same order with one important exception: interest. The MAT's were rated lower in ability to stimulate interest than in any other area except control. Bryan's sample of first year teachers were rated relatively high on this question. Either the MAT's are less able to stimulate interest in their students or their students are harder to stimulate.

The rank order of the questions for pre-test groups are given in Table 1. There were no significant changes in rank order for post-test

---

[15]Roy C. Bryan, Why High School Teachers Use Image Reports (Kalamazoo: Student Reaction Center, 1965), p. 20.

groups combined or for groups taken individually. (The minimum r_s was .90.) Nor were there significant differences between the rank order on the pupil questionnaire and the rank order on the self-appraisal questionnaire. (The minimum r_s was .78.) It is interesting to notice that on questions having to do with enthusiasm or personal acceptance and understanding of pupils the teachers were ranked high, while on questions having to do with discipline, structure and organization they were ranked low. Apparently these beginning teachers were trying to be more of a friend to their students than a parent.

TABLE 1

Rank Order of Ratings on the Pupil Questionnaire

All Pre-test Groups

QUESTION		RANKING
No.	Content	
7	Attitude to Subject	1
8	Attitude to Student Opinion	2
11	Sense of Humor	3
10	Student Participation	4
5	Attitude Toward Students	5
1	Knowledge	6
3	Fairness	7
2	Explanations	8
12	Planning	9
9	Variety	10
6	Interest	11
4	Control	12

On average the teachers rated themselves at about the same level, or slightly lower, than they were rated by their pupils. The only exceptions to this are on clarity of explanations and on planning. In both areas MAT's tended to think worse of themselves than did their students. The differences, although not large, could indicate a tendency for beginning teachers to judge themselves more strictly in these areas than in others. A poorly planned lesson may be more obvious to a teacher than to his students.

Standard Deviations--The standard deviations reported in Tables 8 and 10 indicate the amount of agreement, or disagreement, which classes expressed about teachers. Similarly, the standard deviations in Tables 9 and 11 indicate variability in self-perceptions. According to their pupils, our interns varied most in control on the initial testing. On the second testing, as far as students were concerned, there was slightly more variability between teachers on all items, but control was still the

FIGURE 1

Profile For Pre-test Groups Combined

SCALE STEPS	SCALE AVERAGE	QUESTIONS 1	2	3	4	5	6	7	8	9	10	11	12
Very Best	5.0												
	4.9												
	4.8												
	4.7												
	4.6												
	4.5												
	4.4												
	4.3												
	4.2												
	4.1												
Very Good	4.0												
	3.9												
	3.8												
	3.7												
	3.6	B, A											
	3.5												
	3.4												
	3.3												
	3.2												
	3.1												
Good	3.0												
	2.9												
	2.8												A
	2.7												
	2.6												
	2.5												
	2.4												
	2.3												B
	2.2												
	2.1												
Average	2.0												
	1.9												
	1.8												
	1.7												
	1.6												
	1.5												
	1.4												
	1.3												
	1.2												
Below	1.1												
Average	1.0												

A Pupil Questionnaire B Self-Appraisal Questionnaire

KEY TO QUESTIONS

1. Knowledge
2. Explanations
3. Fairness
4. Control
5. Attitude toward students
6. Interest
7. Attitude toward subject
8. Attitude toward student opinions
9. Variety
10. Student participation
11. Sense of humor
12. Planning

most variable. There was also more variability on the self-appraisal questionnaire at the second testing, but interest and sense of humor were as variable as control.

The standard deviations in Tables 8-11 do not tell us anything about the amount of agreement or disagreement within classes. As we shall see later in this report, the within class variation is considerably higher than the between class (or teacher) variation.

Skewness and Kurtosis--Some of the subsequent tests of differences between groups assume that the distribution of scores within groups is a normal distribution (multivariate normal in the case of the multivariate tests). The third and fourth columns of Tables 8-11 give us an indication of the validity of this assumption. In general, there seems to be no reason to question it. The data for groups taken separately are essentially the same as in Tables 8-11.

CORRELATION AND FACTOR ANALYSIS

The correlation matrix for class means on the pupil questionnaire for all pre-test groups combined are presented in Appendix E, Table 12. As can be seen the questions are highly correlated with many of the correlation coefficients significant beyond the .01 level. Notice that these are correlations between class means not within class correlations.

A factor analysis of this correlation matrix yields essentially one factor. A second factor can be extracted, but it has a latent root of only .082. The factor loadings are given in Appendix E, Table 13. It indicates that the pupil questionnaire has a single dimension without much dispersal along this dimension. Control, and to a lesser extent planning are towards one end of this dimension with the rest of the questions clustered together. One is tempted to speculate what lies at the extremes of this dimension and what other dimensions there are to pupils' conceptions of their teachers.

The correlation matrix for the self-appraisal questionnaire for all pre-test groups combined is presented in Appendix E, Table 14. These items are also highly correlated, but apparently not so highly as are the items on the pupil questionnaire. This is most probably due to the fact that the measures on the self-appraisal questionnaire are discrete variables.

ANALYSIS OF GROUP DIFFERENCES USING A TOTAL GAIN SCORE

This analysis was intended to detect differences among the three groups that were given the pupil questionnaire both in February and in May without giving consideration to the differential effects due to the different items. A single total gain score was computed by adding together all 12 May ratings and subtracting from this the sum of the February ratings. Three conclusions can be drawn:

1. The grand mean of the total gain scores was less than zero (-.626), but it was not significantly different from zero ($F = 1.8714$ with 1 and 42 degrees of freedom). Thus these data give us no reason to think that there was an overall tendency to gain or to lose during the experimental period.

2. The means for the two experimental groups were not significantly different ($F = .0772$ with 1 and 42 df). Therefore, these two groups were pooled together and compared with the control group.

3. There was a significant difference between the control group and the experimental groups ($F = 4.54$ with 1 and 42 df; $p < .04$). Thus the basic hypothesis of this study was supported: feedback was effective in producing changes in perceived behavior. But when we looked at the estimates of effects, we found that the changes were in the opposite direction than we expected. Feedback was effective in lowering teachers' total gain score. The mean total change score for the three groups are given in Table 2.

TABLE 2

Mean Total Change Scores for Three Treatment Groups

Control	Experimental I	Experimental II
.745	-1.16	-1.47

As we have remarked in the previous section, the process of computing a total gain score resulted in a loss of information. What was needed was a method of analysis which could, in effect, unpool the individual differences between items; i.e. a multivariate approach to the data.

ANALYSIS OF GROUP DIFFERENCES USING MULTIVARIATE ANALYSIS OF VARIANCE

This analysis was intended to uncover differences between the three groups which the previous analysis did not detect. Given that feedback had a negative effect, to which items could this effect be attributed? Twelve gain scores were computed by subtracting the February rating on an item from the corresponding May rating. The analysis pointed to three areas in which interesting conclusions can be drawn:

1. The grand mean vector of the gain scores was significantly different from zero ($F = 6.6476$ with 12 and 31 df; $p < .0001$). This appears to differ from the previous analysis; but it can be explained by the fact that, overall, significant gains in control and variety are

balanced by significant losses in attitude toward student, attitude toward student opinion, and fairness. The drop in fairness is particularly marked as is the gain in variety.

2. There is a highly significant difference between the mean vectors of the three groups ($F = 2.7230$ with 24 and 62 df; $p < .0009$). Thus, the multivariate test shows that there are important differential effects between the items.

3. The difference between the two experimental groups is marked ($F = 4.8236$ with 12 and 31 df; $p < .0003$). This again appears to differ from the previous results, but again some losses are balanced by gains. The crucial items appear to be: knowledge of subject on which Experimental Group I showed a gain while Experimental Group II showed a loss; planning on which Experimental Group I showed a loss while Experimental Group II showed a gain; and attitude toward students on which both groups showed a loss but Experimental Group II showed an extremely large loss. Control also seems to distinguish these groups, Experimental Group I showed a loss while Experimental Group II showed a gain, but the effect is not as significant. It would seem that the conference had an effect in improving aspects of management and discipline with a consequent drop in perceived attitude toward students and knowledge of subject matter. The mean vectors of the gain scores for the three groups are in Table 3.

TABLE 3

Mean Vectors of Gain Scores for Three Treatments

ITEM	CONTROL	EXP. I	EXP. II
1. Knowledge	.114	.060	-.177
2. Explanations	.153	-.141	.070
3. Fairness	-.160	-.127	-.258
4. Control	.173	-.047	.175
5. Attitude Toward Students	.015	-.081	-.425
6. Interest	.103	-.039	-.201
7. Attitude to Subject	-.138	-.138	-.167
8. Attitude to Student Opinion	-.088	-.146	-.251
9. Variety	.368	.023	.012
10. Student Participation	.074	-.137	-.071
11. Sense of Humor	.030	-.134	-.198
12. Planning	.010	-.250	.025

ANALYSIS OF GROUP DIFFERENCES USING T-TESTS

The preceding analyses have shown that there were significant differences between the treatments. Yet it is hard to untangle these effects. Another way to look at the data is to consider them class by class and item by item. This was done using t-tests of the significance of a difference between a rating in May and a rating in February.

A first look at the summary of t-tests (see Appendix E, Table 15) shows that there are a number of significant gains (or losses). In fact, 55 of these are significant at the .05 level or better. One must remember that out of the total of 540 items tested one would expect in the neighborhood of 5 percent, or 27, to be significant even if the null hypothesis were in fact true. The question is whether the pattern of significant gains or losses is sufficiently different from what one would expect from purely chance expectation. The data are laid out in Table 4. Chi-square was computed for each group separately. The conclusion is clear; the pattern of significant gains and losses is close to what one would expect for the control group, but not close in the experimental groups.

TABLE 4

Chi-square Analysis of the Pattern of Significant Gains and Losses

		p>.10		.10>p>.05		.05>p>.02		.02>p			
	N	E	A	E	A	E	A	E	A	χ^2	P
Control Group	180	162	157	9	10	5.4	8	3.6	5	2.06	≈.55
Experimental Group I	192	172.8	156	9.6	11	5.76	14	3.84	11	25.4	<.001
Experimental Group II	168	151.2	143	8.4	8	5.04	8	3.36	9	11.6	<.02

E = expected frequency under the null hypothesis
O = observed frequency

However, this analysis does not tell us about the direction of change. To answer this question one can view the data by considering the total number of gains and losses in each group. If there were no systematic change in pupils' evaluation of their teachers during the experimental period, we would expect the number of gains and losses in the control group to be approximately the same. In fact, there are 105 gains and 75 losses out of 180 possibilities (not including question 13). Is this evidence of some systematic tendency towards improved scores? Using the sign-test we can test the null hypothesis that the median of the differences is zero against the alternative that the median is non-zero. For 105 gains out of a sample of 180 the p-value is .031. This is evidence that without feedback there was an overall tendency for scores to gain during our experimental period. If one considers only the significant gains and losses in the control group, there were 7 gains and 6 losses, a fact which neither supports nor detracts from the conclusion. Yet it is interesting to note that 5 out of the 6 significant losses were achieved by a single teacher.

If we look at Experimental Group I (those who received written feedback) in the same way, we arrive at a surprising result. There were only 74 gains out of a possible 192 for which the p-value is .0019. Evidently with written feedback there was an overall tendency for scores to lose during the experimental period. This evidence is given some added weight when one notices that 17 of the losses were significant (at the .05 level or better) while only 8 of the gains were significant. Testing the null hypothesis that the median significant difference is zero against the alternative that it is negative, we get a p-value of .054.

Looking at Experimental Group II there were 67 gains and 101 losses yielding a p-value of .011. Thus, with written feedback and a conference there was also an overall tendency for scores to drop during the experimental period. Again if one looks only at the significant gains and losses there were 18 losses and 1 gain which yields a p-value of .0000382.

There is, however, some further information with this group. Each teacher chose two items for improvement. None of the selected items showed a gain significant beyond the .05 level, but there were 19 gains and 8 losses on these items. Testing the null hypothesis that the median of these differences was zero against the alternative that it was negative, the p-value is .025. Apparently the conference and the selection of two areas for improvement counteracted the overall negative influence of the written feedback.

To summarize the preceeding analysis of the data:

1. Those teachers who received no feedback showed a highly significant tendency to gain during the experimental period, but the gains were not significant.

2. Those teachers who received written feedback showed a highly significant tendency to lose during the experimental period, and the losses were somewhat significant.

3. Those teachers who received written feedback and a conference showed a highly significant tendency to lose during the experimental period except on the item selected for improvement, where there was a significant tendency to improve. For these teachers, the overall tendency to lose was somewhat significant, but the gains on selected items were not significant.

4. In each group there was one "loser". Elimination of these three from the sample would not change the above conclusions.

MULTIVARIATE ANALYSIS TO DETECT POSSIBLE SOURCES OF INVALIDITY

Two groups were included in the study in order to give us a check on some possible sources of invalidity. The post-test only group was used to give an indication of whether there was any tendency for ratings

to change due to history, maturation, or some other effect. The self-appraisal group was used to see if the self-appraisal by itself could produce any change in behavior. Two conclusions can be drawn:

1. There were no significant differences between the three groups tested in February. Nor would we expect there to be any since teachers were randomly assigned to these groups.

2. There were no significant differences between the three groups tested in February, the self-appraisal group, and the post-test only group. Thus there is no evidence of any tendency for pupils to change their ratings between February and May for reasons others than the experimental treatments. Nor is there any evidence of any reactive effects due to the administration of the questionnaires.

A word of caution is in order here, particularly since these results contradict some earlier studies. The post-test only group and the self-appraisal group were smaller than the others because we were less interested in drawing conclusions from them. Thus the fact that we saw no general tendency for scores to change during the experimental period may only indicate that we did not look hard enough for it.

WITHIN-GROUP STATISTICS

The within-group statistics were computed in order to gain an impression of the amount of disagreement within a class. All of the comparisons between groups as well as the correlation and factor analyses were done with class means. The within-group statistics give us an idea of how well the class means represent the feelings of the whole class.

The results for the pre-test groups combined are shown in Table 5. They indicate that there is a wide disagreement on all items, even on the homework item. This mean deviation in most cases is about one scale unit. Thus one standard deviation on either side of the mean takes up one-half of the scale. The pattern for other groups is similar.

TABLE 5

Mean Deviation and Total Number of Students Tested
All Pre-Test Groups (48 classes)

ITEM	MEAN DEVIATION	N
1. Knowledge	0.873	1115
2. Explanations	1.09	1108
3. Fairness	1.18	1105
4. Control	1.03	1116
5. Attitude Toward Students	1.09	1107
6. Interest	1.20	1108
7. Attitude to Subject	1.01	1109
8. Attitude to Student Opinion	1.08	1108
9. Variety	1.18	1109
10. Student Participation	1.15	1112
11. Sense of Humor	1.09	1103
12. Planning	1.13	1098
13. Homework	0.714	1085
Average number of students per class:	23.3	

ACCURACY OF SELF-APPRAISAL

Correlation coefficients between the self-appraisal scores and the corresponding class means on the pupil questionnaire are given in Table 6. They show that the combined pre-test groups were accurate in predicting their pupils' responses on control, variety, student participation, sense of humor, and planning. To a lesser extent they predicted the response for knowledge of subject matter, but their predictions were no better than chance in the other areas.

In a sense, the items on which interns were accurate in their appraisal are the ones for which there is an objective answer. The other items are the ones for which we have to accept a student's opinions. If he tells us that his teacher's explanations are not clear to him, we have to accept that. Similarly, he is the best judge of whether or not a class is interesting. The implication seems to be that teachers are able to judge things from a standard frame of reference, but they are not able to accurately see things from their pupils' frame of reference.

TABLE 6

Correlations Between Self-Appraisal and Pupil Questionnaire
All Pre-Test Groups

ITEM	CORRELATION COEFFECIENT	PERCENTAGE OF VARIANCE ACCOUNTED FOR
1. Knowledge	.306*	.094
2. Explanations	.081	.007
3. Fairness	.050	.002
4. Control	.506**	.257
5. Attitude Toward Students	.218	.048
6. Interest	.227	.052
7. Attitude to Subject	.270	.073
8. Attitude to Student Opinion	.275	.076
9. Variety	.460**	.212
10. Student Participation	.498**	.248
11. Sense of Humor	.500**	.250
12. Planning	.403**	.162

* $.01 < p < .05$ ** $p < .01$

Residual variances were compared to see if there were any changes in accuracy of self-appraisal due to feedback. If the residual variances went up between February and May, it was considered to be a loss in accuracy. Only one of these gains or losses was significant beyond the .05 level. Yet when one considers the pattern of gains and losses which are shown in Table 7, one is struck by the fact that Experimental Group

II made a large number of gains. Putting it the other way around, there were 8 losses out of 12 for the Control Group, 7 out of 12 for Experimental Group I, and only 2 out of 12 for Experimental Group II. Using a sign test, the probability that this would happen by chance is less than .025 for Experimental Group II, but it is not significant for the other groups. Evidently the conference was effective in focusing teachers' attention on the meaning of the feedback.

TABLE 7

Gains and Losses in Accuracy of Self-Appraisal for Three Treatments

ITEM	CONTROL	EXP. I	EXP. II
1. Knowledge	-	+	+ *
2. Explanations	-	-	+
3. Fairness	+	-	+
4. Control	-	-	+
5. Attitude Toward Students	+	+	+
6. Interest	-	+	+
7. Attitude to Subject	-	-	+
8. Attitude to Student Opinion	+	-	-
9. Variety	-	-	+
10. Student Participation	+	-	-
11. Sense of Humor	-	+	+
12. Planning	-	+	+
No. of losses	8	7	2

- denotes a loss in accuracy from February to May

+ denotes a gain in accuracy from February to May

* $p < .05$

DISCUSSION

The underlying hypothesis of this study was that feedback from the students would be helpful in improving the performance of beginning teachers. Not only was this not substantiated, we observed exactly the opposite effect. The data are clear: feedback was effective in lowering students' ratings of their teachers.

There are two possible kinds of explanation for this surprising result. The first is that the drop is due primarily to characteristics of the student raters rather than to a change in the behavior of the teacher. In micro-teaching situations, for example, it has been found that student raters tend to become hypercritical before they become sophisticated. They are generally complimentary for the first few ratings, but soon abandon this stance to become quite harsh in their judgments. Only after this do they begin to make distinctions between different teaching performances. If the same process occurrs with feedback instruments administered in a regular classroom situation, we would expect it to have the same effect on all groups. In particular, we would expect the teachers in the control group to be rated lower in May, but, in fact they tended to be judged less harshly. Perhaps constant exposure to a teacher already had made his students more discriminating about his characteristic performance.

One could also argue that the fact that a teacher received feedback made his students more willing to be frank for the second rating. If students fear the sanctions their teacher might impose were they to tell him their real feelings and if they know that their teacher has received some feedback from them, then the fact that no sanctions were forthcoming might encourage them to be more straightforward. Or perhaps they do not fear the sanctions, but when they see no change in their teacher's behavior they express their opinions more bluntly. Both of these speculations assume that the students of teachers in the control group knew, or suspected, that their teachers had not received any feedback at the time of the second rating.

A more plausible kind of explanation of the results is that feedback did, in fact, change teachers' behavior. On this view, receiving a summary of their students' opinions had a disorienting effect on the teachers. Although they valued the source of the feedback, they did not know how to use the information which it provided. It might have encouraged them to abandon methods with which they were comfortable and to substitute different and unfamiliar procedures which they had not carefully considered.

This explanation is given added weight when one remembers that there were significant differences between the group which received written

feedback only and the group which received written feedback plus a conference. The conference group was rated higher in May on aspects of control and planning, where they were initially lowest, but considerably lower in their attitude toward students, where they were initially highest. They also tended to be rated higher in May on the items selected for improvement. These results indicate that the conference was helpful in overcoming the disorienting effect of the feedback by providing direction in classroom management. This had the effect, however, of making the teacher appear to have a less positive attitude toward his students.

The conference also seems to have had a significant effect in improving the accuracy of teachers' self-appraisal. Although they may not have been able to improve their ratings, they had a clearer idea of where they stood. This indicates that teachers who received written feedback only did not really understand the message. Because it was threatening or otherwise disorienting, the information in the written feedback was ignored or discounted. Nevertheless the shock of receiving it resulted in a drop in performance. The conference focused teachers' attention on the real meaning of the feedback, and perhaps directed their attention to some of its more positive aspects.

Our results are in conflict with earlier findings in two other respects, both relatively minor. First, we noted no tendency for scores to change between February and May. Students do not seem to express harsher opinions later in the year, unless their teachers have received feedback. If we had held the second rating in June, just before teachers gave out grades, perhaps we would have noticed such a tendency. As we have pointed out, our experiment was not designed to place maximum emphasis on detecting this effect. Perhaps a future experiment might systematically trace the ups and downs of student opinion during the course of a year.

We also did not notice any tendency for our instruments to have any effect on subsequent ratings. We were interested to see if the self-appraisal, by itself, would have focused teachers' attention on their strengths and weaknesses with a consequent rise in performance; but all groups were so inaccurate in their self-appraisals that it became unlikely that this instrument alone would have any effect on later pupil ratings.

The results which we have presented contain at least three kinds of implications for future work in this area. First, the feedback instrument needs to be revised and extended. Several directions which this could take are indicated.:

(a) A thorough attempt needs to be made to discover the structure of students' thinking about teachers. The factor analysis showed that all the items on the questionnaire were measuring roughly the same thing when averaged over a class. Thus, the questionnaire taps only one dimension, and a limited part of that. What are other dimensions? What would students like to tell us about their teachers?

(b) As well as class means, the feedback should show clusters of disagreements within a class. There is considerable difference of opinion

within a class about the teacher. In this situation, a class mean does not provide us with the most interesting information. A teacher would like to know more about the variety of opinion which students have about him.

(c) Some consideration should be given to the question of scaling. If the range of the scale were larger, it might not result in a comparable increase in variance within a class. On the other hand, if there were too many intervals, the full range of the scale might not be used or there might be a tendency to respond towards one end of the scale. The optimum number of intervals is a complicated, yet important question.

Another implication of these results is that more attention needs to be given to developing the most effective way of administering the feedback. Written feedback is too damaging, at least for beginning teachers, and the message does not get through without being garbled. The conference mitigated this effect, but it was not as helpful as we had wished. If we persist in believing, as we do, that feedback can be helpful, how can we provide it in a less damaging and more informative manner?

A third implication of these results is that the effects of feedback need to be traced over a longer interval than eight weeks of school. It seems likely that feedback may result in a sharp initial drop in performance which is then recovered as the teacher learns to assimilate and respond to the new situation. From this point of view we would expect a drop in ratings over a short period of time but a gain in ratings over a longer period of time. This pattern, if it exists, would make our results more consistent with earlier findings.

One way to check this hypothesis would be to administer feedback instruments over the course of a school year at, say, four week intervals. By varying the ways in which feedback was administered and by tracing the pattern of response to these treatments, we would gain a more accurate picture of the dynamics of using student feedback with beginning teachers.

REFERENCES

Aubertine, H. E. "An Experiment in the Set Induction Process and Its Application in Teaching." Unpublished Ph.D. dissertation, Stanford University, 1966.

Bock, R. D. "Programming Univariate and Multivariate Analysis of Variance." Technometrics, V (February, 1963), 95-116.

Bock, R. D., and Haggard, E. A. "The Use of Multivariate Analysis of Variance in Behavioral Research." In Dean Whitla (Ed.) Handbook of Measurement in Education, Psychology, and Sociology. Boston: Addison-Wesley, 1968.

Brownlee, K. A. Statistical Theory and Methodology In Science and Engineering. New York: John Wiley & Sons, Inc., 1960.

Bryan, Roy C. Reactions to Teachers by Students, Parents and Administrators. Report of Cooperative Research Project, No. 668, U. S. Office of Education. Kalamazoo, Michigan: Western Michigan University, 1963.

Bryan, Roy C. Why High School Teachers Use Image Reports. Kalamazoo, Michigan: Student Reaction Center, 1965.

Finn, J. D. "Univariate and Multivariate Analysis of Variance and Covariance." Research Memorandum No. 3, Statistical Laboratory, Department of Education, The University of Chicago, April, 1966.

Gage, N. L., Leavitt, G. S., and Stone, George C. "Teachers' Understanding of Their Pupils and Pupils' Rating of Their Teachers." Psychological Monographs, LXIV, No. 21 (Whole No. 406), 1955.

Gage, N. L., Runkel, P. J., and Chatterjee, B. B. Equilibrium Theory and Behavior Change: An Experiment in Feedback from Pupils to Teachers. Urbana, Illinois: University of Illinois Bureau of Educational Research, 1960.

Gage, N. L., Runkel, P. J., and Chaterjee, B. B. "Changing Teacher Behavior Through Feedback from Pupils: An Application of Equilibrium Theory." Readings in the Social Psychology of Education. Edited by W. W. Charters and N. L. Gage. Boston: Allyn and Bacon, 1963.

Remmers, H. H. "Rating Methods in Research on Teaching." Handbook of Research on Teaching. Edited by N. L. Gage. Chicago: Rand McNally and Company, 1963.

Rulon, P. J., and Brooks, W. D. "On Statistical Tests of Group Differences." In Dean Whitla (Ed.) Handbook of Measurement in Education, Psychology, and Sociology. Boston: Addison-Wesley, 1968.

Ryan, Kevin A. "The Use of Students' Written Feedback in Changing the Behavior of Beginning Secondary School Teachers." Unpublished Ph.D. dissertation, Stanford University, 1966.

Tuckman, B. W., and Oliver, W. F. "Effectiveness of Feedback to Teachers as a Function of Source." Journal of Educational Psychology, LIV (1968), 297-301.

APPENDIXES

APPENDIX A

STUDENT-OPINION QUESTIONNAIRE

Please answer the following questions honestly and frankly. Do not give your name. To encourage you to be frank, your regular teacher should be absent from the classroom while these questions are being answered. Neither your teacher nor anyone else at your school will ever see your answers.

The person who is temporarily in charge of your class will, during this period, collect all reports and seal them in an envelope. Your teacher will receive from The University of Chicago a summary of the answers by the students in your class. The University will not mail this summary to anyone except your teacher unless requested to do so by your teacher.

After completing this report, sit quietly or study until all students have completed their reports. There should be no talking.

Use the answer sheet provided to indicate your answers to Questions 1 - 13. Answer Questions 14 and 15 on this sheet.

WHAT IS YOUR OPINION CONCERNING THIS TEACHER'S:

1. KNOWLEDGE OF SUBJECT: Does he have a thorough knowledge and understanding of his teaching field?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

2. CLARITY OF EXPLANATIONS: Are assignments and explanations clear?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

3. FAIRNESS: Is he fair and impartial in his treatment of all students?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

4. CONTROL: Does he keep enough order in the classroom? Do students behave well?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

5. ATTITUDE TOWARD STUDENTS: Is he patient, understanding, considerate, and courteous?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

6. ABILITY TO STIMULATE INTEREST: Is this class interesting and challenging?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

7. ATTITUDE TOWARD SUBJECT: Does he show interest in and enthusiasm for the subject? Does he appear to enjoy teaching this subject?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

8. ATTITUDE TOWARD STUDENT OPINIONS: Are the ideas and opinions of students treated with respect? Are differences of opinion welcomed even when a student disagrees with the teacher?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

9. VARIETY IN TEACHING PROCEDURES: Is much the same procedure used day after day and month after month, or are different and appropriate teaching methods used at different times (student reports, class discussions, small-group discussions, films and other audio-visual aids, demonstrations, debates, field trips, teacher lectures, guest lectures, etc.)?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

10. ENCOURAGEMENT OF STUDENT PARTICIPATION: Do students feel free to raise questions and express opinions? Are students encouraged to take part?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

11. SENSE OF HUMOR: Does he see and share with students amusing happenings and experiences?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

12. PLANNING AND PREPARATION: Are plans well made? Is class time well spent? Is little time wasted?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

13. ASSIGNMENTS: Are assignments (out-of-class, required work) sufficiently challenging without being unreasonably long? Is the weight of assignments reasonable?
a) Much too light b) Too light c) Reasonable d) Too heavy e) Much too heavy

14. Please name two or more things that you especially like about this teacher or course.

15. Please give two or more suggestions for the improvement of this teacher or course.

Reproduced by permission of:

Student Reaction Center
Western Michigan University
Kalamazoo, Michigan

APPENDIX B

SELF-APPRAISAL QUESTIONNAIRE

Please answer the following questions honestly and frankly. Although you are asked to give your name, the results will be held in strict confidence. They will never be used as part of an evaluation of your teaching performance. You will receive a summary of how the students in your class have answered the same questions.

Use the answer sheet provided to record your answers to Questions 1 - 13. You should answer the questions in the way in which you anticipate that your students would answer the same questions. We are interested in the way in which you think they perceive you, not in the way in which you perceive yourself.

1. KNOWLEDGE OF SUBJECT: Does he have a thorough knowledge and understanding of his teaching field?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

2. CLARITY OF EXPLANATIONS: Are assignments and explanations clear?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

3. FAIRNESS: Is he fair and impartial in his treatment of all students?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

4. CONTROL: Does he keep enough order in the classroom? Do students behave well?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

5. ATTITUDE TOWARD STUDENTS: Is he patient, understanding, considerate, and courteous?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

6. ABILITY TO STIMULATE INTEREST: Is this class interesting and challenging?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

7. ATTITUDE TOWARD SUBJECT: Does he show interest in and enthusiasm for the subject? Does he appear to enjoy teaching his subject?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

8. ATTITUDE TOWARD STUDENT OPINIONS: Are the ideas and opinions of students treated with respect? Are differences of opinion welcomed even when a student disagrees with the teacher?
   a) Below Average b) Average c) Good d) Very Good e) The Very Best

9. VARIETY IN TEACHING PROCEDURES: Is much the same procedure used day after day and month after month, or are different and appropriate teaching methods used at different times (student debates, field trips, teacher lectures, guest lectures, etc.)?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

10. ENCOURAGEMENT OF STUDENT PARTICIPATION: Do students feel free to raise questions and express opinions? Are students encouraged to take part?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

11. SENSE OF HUMOR: Does he see and share with students amusing happenings and experiences?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

12. PLANNING AND PREPARATION: Are plans well made? Is class time well spent? Is little time wasted?
a) Below Average b) Average c) Good d) Very Good e) The Very Best

13. ASSIGNMENTS: Are assignments (out-of-class, required work) sufficiently challenging without being unreasonably long? Is the weight of assignments reasonable?
a) Much too light b) Too light c) Reasonable d) Too heavy e) Much too heavy

APPENDIX C

SAMPLE OF WRITTEN FEEDBACK

THE UNIVERSITY OF CHICAGO
MASTER OF ARTS IN TEACHING PROGRAM

Enclosed is a compilation of student responses to the questionnaire given to one of your classes recently.

On page 1 you will find two graphs. Graph A is an average of student responses to Questions 1 - 12; Graph B, in blue, is your prediction of student responses. Note what your students perceive to be your strengths and weaknesses, paying particular attention to those items where there is considerable disparity between student perceptions and your perceptions. Hopefully this information will suggest areas in which you may wish to make a special effort to change your students' perceptions.

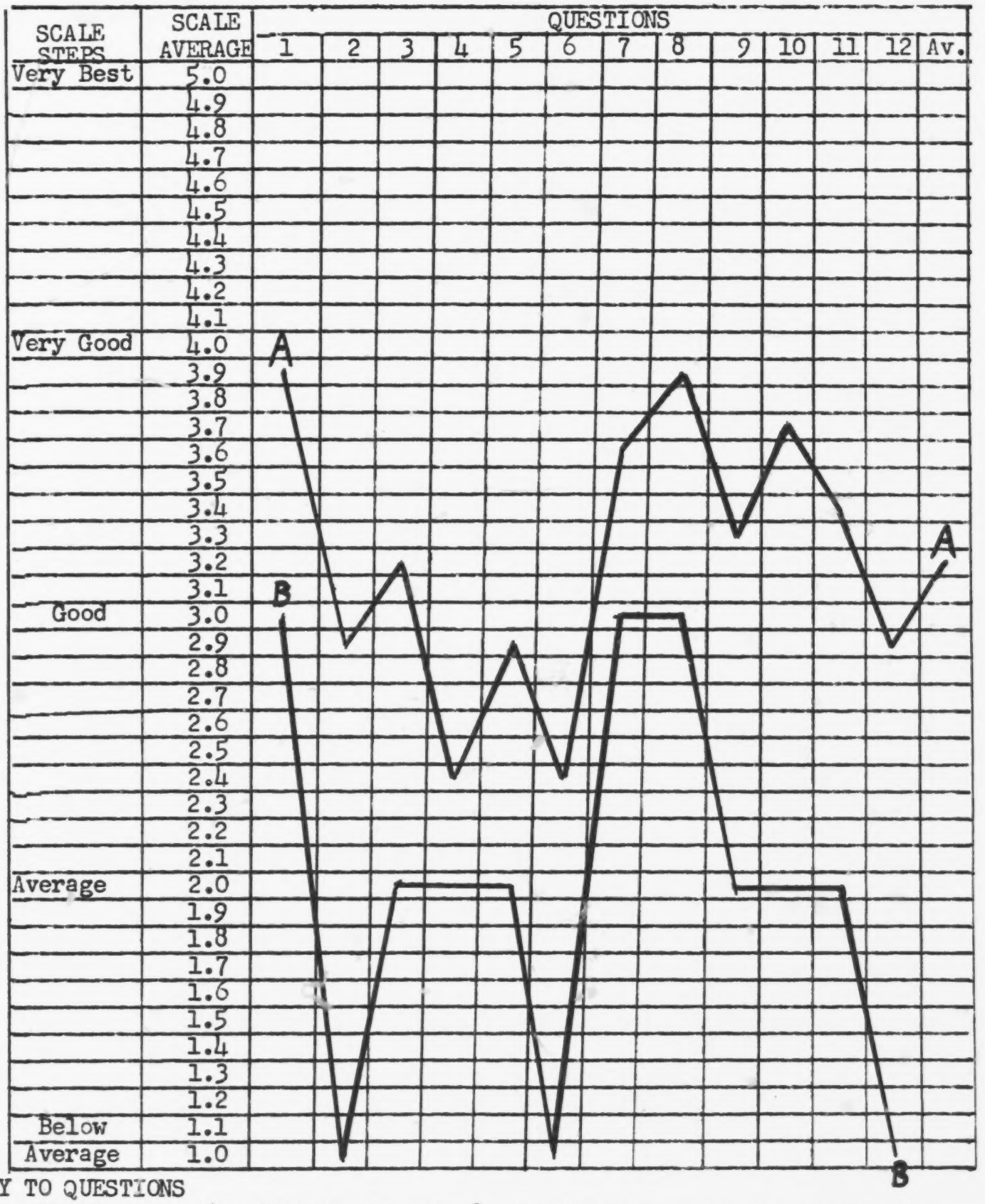
On page 2 is a summary of student comments in response to Questions 14 and 15 on the questionnaire. These comments have been edited only to avoid unnecessary repetition and eliminate irrelevance. Statements which represent frequently mentioned sentiments are followed by an "(F)".

At a later date arrangements will be made with you for a second visit to your class at the end of April.

APPENDIX C (continued)
SAMPLE OF WRITTEN FEEDBACK

Teacher No. 109 February 1968

A —— Pupil Questionnaire B —— Self-Evaluation



KEY TO QUESTIONS

1. Knowledge
2. Explanations
3. Fairness
4. Control
5. Attitude toward students
6. Interest
7. Attitude toward subject
8. Attitude toward student opinions
9. Variety
10. Student participation
11. Sense of humor
12. Planning

Av. = Mean of averages 1-12

APPENDIX C (continued)

SAMPLE OF WRITTEN FEEDBACK

13. HOW STUDENTS ANSWERED THE QUESTION ON WEIGHT OF ASSIGNMENTS:

Much too light_____ Too light_____ Reasonable 22

Too heavy 3 Much too heavy 5

Total number of students 31

SUMMARY OF COMMENTS BY STUDENTS

Teacher 109 February 1968

14. THINGS STUDENTS ESPECIALLY LIKE ABOUT THIS TEACHER OR COURSE:

"...course is very interesting to study (F)...very fair (F); knows what's happening...interesting discussions (F); gives the right amount of homework...he takes a lot off of the kids and yet is great to them. He never really gets mad...sense of humor...cheerful and considerate... doesn't stick to the same thing or routine day after day...he knows how to take a joke but at the same time he's stern, but that's good; he'll give you a second chance if he sees it possible, and if you show some interest...he does have an understanding of "teenage" problems and tries to help us and help us help others; he is very open minded at class discussions about politics and other things...he gives you a passing grade when he's sure you're doing the very best you can..."

15. SUGGESTIONS FOR IMPROVEMENT:

"...plans things but we never finish them (F)...could be more deliberate in classroom discussions and get to the point more quickly and save a little time...make his lectures a little shorter in order for us to be able to ask more questions...gives too many citizenships and doesn't use them like they should be used...I think he should figure out different ways to keep the class quiet instead of quizzes...he should try to know the students more because that's one reason kids don't like teachers...too many tests...he should have more class-group discussions, debates, guest lectures...should go on field trips; should have things for extra credit...he's not fair at all. If he likes you fine, but if he doesn't forget it. You lose in his course...don't give surprise quizzes as punishment...he should also have a set disciplinary policy instead of sending 1 kid down almost every day and another who is equally as loud or rowdy is hardly ever sent..."

APPENDIX D

PROTOCOL FOR FEEDBACK CONFERENCES
(Experimental Group II)

Pleasant greeting and reassuring remarks to put conferee at ease.

Introduction of feedback from students (summary of ratings and written comments) and explanation of its organization, as well as its promise and limitations.

Suggestion that conferee review materials by himself.

Supervisor to leave room and return in about four minutes.

Discussion begun by encouraging conferee to comment on what he understands his students' view of his teaching to be.

Supervisor to reinforce positive aspects of feedback before alluding to areas where need for improvement is indicated.

Supervisor to probe for items on which conferee was surprised, that is, the feedback does not conform to his self-appraisal.

Supervisor to probe for items which have aroused especial interest on the part of the conferee.

Establishment of two areas for intensive improvement effort.

Conferee to make own selection.

Supervisor to encourage selection of areas where conferee feels confident that he can improve.

Reflection on possible reasons for low ratings in the selected areas.

Exploration of several possible courses of action to improve image in selected areas.

Reiteration of agreement on areas specified for concentration of effort and reminder that feedback will be collected again in eight weeks.

APPENDIX E

TABLE 8

Basic Statistics for Class Means
Pupil Questionnaire

All Pre-Test Groups

ITEM	MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
1. Knowledge	3.578	0.452	-0.450	-0.679
2. Explanations	3.055	0.528	0.181	-0.662
3. Fairness	3.518	0.605	-0.142	-1.030
4. Control	2.411	0.762	0.687*	-0.477
5. Attitude Toward Student	3.674	0.659	-0.339	-0.692
6. Interest	2.728	0.557	-0.079	-0.807
7. Attitude to Subject	3.920	0.434	-0.175	-1.104
8. Attit. to Student Opinion	3.894	0.480	-0.545	-0.826
9. Variety	2.801	0.553	-0.227	-0.731
10. Student Participation	3.755	0.482	0.082	-0.058
11. Sense of Humor	3.807	0.615	-0.280	-0.732
12. Planning	2.886	0.601	0.175	0.636
13. Homework	3.087	0.268	-0.264	0.156

N=48 * .01<p<.05

TABLE 9

Basic Statistics
Self-Appraisal Questionnaire

All Pre-Test Groups

ITEM	MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
1. Knowledge	3.562	0.848	-1.360**	1.822**
2. Explanations	2.458	0.874	0.322	0.285
3. Fairness	3.292	0.874	-0.020	-0.840
4. Control	2.333	1.059	0.391	-0.620
5. Attitude Toward Students	3.583	0.871	-0.454	-0.504
6. Interest	2.562	0.796	-0.334	-0.339
7. Attitude to Subject	3.583	0.821	-0.503	0.829
8. Attit. to Student Opinion	3.792	0.798	-0.375	-0.172
9. Variety	2.583	1.088	0.385	-0.289
10. Student Participation	3.729	0.792	-0.264	-0.284
11. Sense of Humor	3.562	0.943	-0.720*	-0.034
12. Planning	2.437	0.897	0.100	-0.719
13. Homework	2.771	0.751	-0.214	-0.217

N=48 * .01<p<.05 ** p<.01

TABLE 10

Basic Statistics for Class Means
Pupil Questionnaire

All Post-test Groups

ITEM		MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
1.	Knowledge	3.662	0.466	-0.189	0.250
2.	Explanations	3.080	0.621	-0.195	-0.426
3.	Fairness	3.363	0.656	-0.133	-0.331
4.	Control	2.477	0.770	0.563	0.242
5.	Attitude Toward Students	3.621	0.635	-0.177	-0.522
6.	Interest	2.713	0.641	0.274	0.291
7.	Attitude to Subject	3.810	0.486	-0.694*	0.645
8.	Attit. to Student Opinion	3.750	0.589	-0.865**	0.930
9.	Variety	2.889	0.634	-0.372	-0.146
10.	Student Participation	3.730	0.521	-0.524	0.186
11.	Sense of Humor	3.693	0.709	-0.707*	0.221
12.	Planning	2.913	0.622	-0.078	0.367
13.	Homework	3.112	0.272	0.006	-0.031

N=66 * .01<p<.05 ** p<.01

TABLE 11

Basic Statistics
Self-Appraisal Questionnaire

All Post-test Groups

ITEM		MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
1.	Knowledge	3.515	0.789	-0.807**	-0.358
2.	Explanations	2.712	0.799	0.196	0.112
3.	Fairness	3.273	1.001	-0.473	-0.539
4.	Control	2.000	0.894	0.650*	-0.266
5.	Attitude Toward Students	3.394	0.875	-0.436	-0.287
6.	Interest	2.758	1.009	0.136	-0.355
7.	Attitude to Subject	3.500	0.809	-0.264	-0.474
8.	Attit. to Student Opinion	3.515	0.899	-0.302	-0.151
9.	Variety	2.424	0.878	0.164	-0.647
10.	Student Participation	3.621	0.718	-0.557	0.088
11.	Sense of Humor	3.409	1.067	-0.412	-0.518
12.	Planning	2.500	0.864	0.144	-0.640
13.	Homework	2.938	0.556	-0.582	1.996

N=66 * .01<p<.05 ** p<.01

TABLE 12

Intercorrelations of Class Means for the Pupil Questionnaire
All Pre-test Groups Combined (N = 48)

Var. No.	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.000	.472**	.450**	.102	.490**	.473**	.607**	.436**	.227	.428**	.470**	.402**	.367*
2		1.000	.637**	.365*	.626**	.541**	.484**	.480**	.171	.559**	.455**	.465**	.082
3			1.000	.405**	.832**	.581**	.629**	.710**	.396**	.586**	.560**	.433**	.170
4				1.000	.247	.498**	.238	.241	.171	.332*	.243	.637**	.046
5					1.000	.483**	.598**	.720**	.359*	.638**	.553**	.315*	.100
6						1.000	.563**	.470**	.559**	.637**	.694**	.558**	.012
7							1.000	.676**	.469**	.524**	.584**	.316*	.058
8								1.000	.462**	.683**	.555**	.213	.290*
9									1.000	.466**	.453**	.171	.020
10										1.000	.699**	.320*	.043
11											1.000	.201	-.059
12												1.000	.280
13													1.000

** p < .01

* .01 < p < .05

KEY TO QUESTIONS

1. Knowledge
2. Explanations
3. Fairness
4. Control
5. Attitude toward students
6. Interest
7. Attitude toward subject
8. Attitude toward student opinions
9. Variety
10. Student Participation
11. Sense of humor
12. Planning
13. Homework

TABLE 13

Principal Components Factor Loadings

All Pre-test Groups

ITEM	FACTOR I	FACTOR II	Communality
1. Knowledge	.995	-.047	.991
2. Explanations	.993	-.003	.986
3. Fairness	.995	-.013	.991
4. Control	.969	.239	.996
5. Attitude Toward Students	.993	-.043	.989
6. Interest	.993	.025	.986
7. Attitude to Subject	.997	-.041	.995
8. Attitude to Student Opinion	.996	-.047	.995
9. Variety	.987	-.051	.977
10. Student Participation	.997	-.030	.994
11. Sense of Humor	.994	-.043	.990
12. Planning	.988	.096	.985
13. Homework	.992	-.037	.985
Latent Roots	12.779	.082	12.861

APPENDIX E (continued)

TABLE 14

Intercorrelations of Ratings on the Self-appraisal Questionnaire
All Pre-test Groups Combined (N = 48)

Var. No.	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.000	.477**	.262	.047	.382**	.372**	.527**	.271	.259	.232	.288*	.453**	.240
2		1.000	.295*	.061	.116	.264	.242	-.165	.004	.091	.042	.444**	.001
3			1.000	.399**	.386**	.309*	.262	.242	.175	.332*	.416**	.377**	.039
4				1.000	.177	.353*	.065	.084	.382**	.161	.298*	.426**	.098
5					1.000	.406**	.526**	.515**	.419**	.573**	.473**	.184	.014
6						1.000	.529**	.255	.522**	.314*	.363*	.452**	.113
7							1.000	.384**	.492**	.346*	.474**	.282	.187
8								1.000	.437**	.582**	.442**	.219	-.010
9									1.000	.237	.316*	.343*	.089
10										1.000	.579**	.260	.072
11											1.000	.332*	.036
12												1.000	.215
13													1.000

* .01 < p < .05 ** p < .01

KEY TO QUESTIONS

1. Knowledge
2. Explanations
3. Fairness
4. Control
5. Attitude toward students
6. Interest
7. Attitude toward subject
8. Attitude toward student opinion
9. Variety
10. Student participation
11. Sense of humor
12. Planning
13. Homework

APPENDIX E (continued)

TABLE 15

Summary of t-test comparisons

CONTROL GROUP

TEACHER NUMBER	QUESTIONS 1	2	3	4	5	6	7	8	9	10	11	12
101					*				-	-		
102					-		-		*	-		
103			-	-		-	* -	-		-		
104	-	* -	** -	-	** -	* -	-	-	-	-	-	* -
105	-	-		**	-		-	-				
106			-				-		*	-	-	
107		-	-	-	-	-	-	-		-	-	
108			-		-		-	-			-	-
109											*	*
110	-	-	-		-		-	-				
111											-	-
112			-	-			-	-				-
113	-		-			-	-			-	-	
115				*	-							
116			-	-	-	-	-	-			-	-
Number of Losses	4	4	9	5	8	5	11	8	2	7	7	5	75

- denotes a loss ** $p < .01$ * $.01 < p < .05$

TABLE 15 (continued)

Summary of t-test comparisons

EXPERIMENTAL GROUP I

TEACHER NUMBER	QUESTIONS												
	1	2	3	4	5	6	7	8	9	10	11	12	
201						-		-		-	* -		
202						*	-		*	*			
203	-	-	-	-	-	** -	-		-	* -	-	* -	
204						**	-				**	-	
205			-	-		-	-	-	-	-		-	
206		** -	-	-	* -	* -	-	-	-	-	-	** -	
207	*	-			-	-		-	-	-	-	-	
208		-		-	-	-	-		-	-	-	-	
209		-	-	-	-		-	-		-	-	-	
210	-	-	-	** -	-	-	-	-	-	-	** -	-	
211		-						-		-		-	
212	-	* -	* -	-	-	-	** -	-	** -	* -	* -	* -	
213	-						-		*	*	-	-	
214			-		-			-		-		-	
215	-	-	-	-	-		-	-	-	-	-		
216		-	-	-	-	-	-	-				-	
Number of Losses	5	10	9	9	10	9	12	11	8	12	10	13	118

- denotes a loss ** $p < .01$ * $.01 < p < .05$

APPENDIX E (continued)

TABLE 15(continued)

Summary of t-test comparisons

EXPERIMENTAL GROUP II

TEACHER NUMBER	QUESTIONS 1	2	3	4	5	6	7	8	9	10	11	12
301		-	* -	◯	** -	-	-	-	◯	-	** -	-
302	-	-	-	◯	* -	◯	-	-		-	-	
303			-	-	◯	-	-	◯			-	
304	◯	◯	-		-	-	-	-	-	-	-	-
305	-	◯	-	-	-		-	-	-			
306	-	◯	-	-	-	-	* -	◯	-		-	-
307	-		-	◯	-		-	-		-	-	◯
308	-	◯	-	◯	** -	-	-	* -	-	-	-	-
310		◯			-	◯						
311	** -	* -	* -		** -	-	* -	** -	◯	* -	* -	◯
313	-	◯		◯		-		-	-		-	-
314				◯	-	◯		-				-
315				◯								◯
316	-	◯	* -	◯	* -	-	-	-	-	-	-	-
Number of Losses	9	6	10	3	12	9	10	11	7	7	10	7	101

- denotes a loss ** $p < .01$ * $.01 < p < .05$ ◯ denotes that this behavior was selected for improvement.